LEARNING MADE EASY

Snowflake Special Edition

Cloud Data Analytics

for dummies®
A Wiley Brand

What is modern cloud data analytics?

What's possible when you unite your siloed data

How to easily obtain data-driven insights

Brought to you by

snowflake™

David Baum

About Snowflake

Snowflake delivers the Data Cloud — a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud.

Snowflake.com.

Cloud Data Analytics

Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Analytics For Dummies®, Snowflake Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-78175-2 (pbk); ISBN 978-1-119-78174-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. Some of the people who helped bring this book to market include the following:

Development Editor: Brian Walls

Project Manager: Martin V. Minner

Senior Managing Editor: Rev Mengle

Acquisitions Editor: Ashley Coffey

Business Development Representative: William Hull

Production Editor: Mohammed Zafar Ali

Snowflake Contributors Team: Jena Donlin, Vincent Morello, Clarke Patterson, Shiyi Gu, Alan Eldridge, Mike Klaczynski, Todd Talkington, Christina Jimenez, Leslie Steere

Table of Contents

INTRODUCTION 1
About This Book 2
Icons Used in This Book 2
Beyond the Book 2

CHAPTER 1: **Defining the Urgency for Modern Cloud Data Analytics** 3
Tracing Analytics History 4
Evolving from Reactive to Predictive Insights 4
Democratizing Analytics 7
Serving all types of users 7
Looking beyond the enterprise 8
Utilizing all types of data 9

CHAPTER 2: **Delivering on the Promise of Analytics** 11
Understanding Data Services 11
Resolving Problems with Siloed Data 13
Engineering data pipelines 14
Accommodating external tables 14
Dealing with streaming data 14
Introducing Data Governance 15
Controlling data access 16
Examining security basics 16
Enforcing consistency with metadata 17
Paying attention to data quality 17
Cataloging data in a schema or catalog 18
Complying with government and industry regulations 18
Adopting Advanced Analytics 20
Embedding Analytics into Applications 20
The Value of a Cloud Data Platform 21

CHAPTER 3: **Unlocking the Power of Your Data** 23
Building Organizational Trust in Cloud Data Analytics 23
Laying a Foundation for Pervasive Business Intelligence 24
Discerning the difference between facts and intuition 25
Moving beyond historical analysis 25

Understanding the Potential of Data Science 26
Improving data science workflows 27
Automating machine learning tasks 29
Writing Effective Queries 30
Leveraging SQL for Multiple Activities 32

CHAPTER 4: **Integrating Cloud Data Analytics into Your Technology Ecosystem** 35
Understanding the Limitations of Traditional Architectures 35
Outlining the Advantages of a Cloud Data Platform 37
Minimizing maintenance 37
Mitigating costs 38
Maximizing performance 38
Utilizing Multiple Clouds 39
Understanding why a cloud data platform is necessary 40
Sharing data efficiently 41
Going Deeper with Visualization 42
Embedding Analytics into Applications 44
Automating Integration Tasks 44
Extracting, Transforming, and Loading Data 46
Standardizing on ANSI SQL 46
Taking Full Advantage of the Cloud 47

CHAPTER 5: **Driving Organizational Transformation and Alignment** 49
Building Data Literacy 49
Gaining Executive Guidance 50
Building a Center of Excellence 50
Stratifying Analytic Users 51
Building a Core Team 52
Sharing Control Between Business and IT 53
Moving Up the Maturity Curve 53
Adhering to Ethical Practices 55
Considering the Future of Analytics 56

CHAPTER 6: **Six Steps for Getting Started with Cloud Data Analytics** 57
Step 1: Review Your Current State 57
Step 2: Enlist Your Team 58
Step 3: Lay a Solid Foundation 59
Step 4: Migrate Data 59
Step 5: Launch a Pilot Project 60
Step 6: Prepare for Growth 60

Introduction

Which trends and developments have the biggest impacts on the success or failure of your organization? Economic, political, and societal challenges emerge often, sometimes overlapping one another. In 2020 alone, we experienced a global pandemic that caused countrywide shutdowns more than once. We witnessed a racial justice movement that began in the U.S. and spread around the world. Other challenges emerged, and even more will come.

How do you prepare your organization to respond quickly to any of these types of impacts in order to remain viable? What about unique market opportunities that remain hidden? For both, how do you deliver new, data-driven insights to reveal where, when, and how your organization should focus over the next 3, 6, or 12 months?

Successful organizations have learned to easily access, unify, integrate, analyze, share, and even monetize data of many types and in exponentially larger amounts. They know how to acquire new data sets that exist outside their domain and easily combine them with their existing data to reveal fresh insights. And they have established a culture of analytic literacy and democracy that informs the important decisions their people make every day. Armed with that information, they can pivot quickly when new events and opportunities impact their organizations at local and global levels.

The cloud has been a boon for these analytic endeavors because it allows organizations to inexpensively store and analyze all the data they need and use that data to detect threats, create new products and services, improve recommendation systems, and otherwise improve business outcomes. Even small companies with limited budgets can take advantage of technologies formerly available only to large organizations with well-funded IT teams.

However, adopting cloud data analytics is not just a matter of repurposing yesterday's on-premises technologies, or moving existing analytic applications and databases from your data center to a cloud vendor's infrastructure. Properly leveraging the power and scale of the cloud requires a new mindset, a new set of management principles, and a new set of cloud-built capabilities.

About This Book

You will find this book important if you want to:

- Supply business users with simple but powerful analytics, without the complexity of managing a data warehouse, data lake, or other types of database management system
- Ensure the security, performance, and reliability of essential analytics processes such as data visualization, data mining, business intelligence, and data science
- Efficiently share and monetize your data to maximize its potential impact, without having to copy or move data

Icons Used in This Book

Throughout this book, the following icons highlight tips, important points to remember, and more:

TIP

Advice about how to maximize analytics in your organization

REMEMBER

Concepts worth remembering as you immerse yourself in understanding today's analytic platforms, processes, and tools

CASE STUDY

Real-world stories about organizations that are using analytics to improve their businesses in innovative ways

TECHNICAL STUFF

The jargon beneath the jargon, explained

Beyond the Book

If you like what you read in this book, visit `www.snowflake.com` to learn more about analytic solutions from Snowflake and its partners, order a free trial of Snowflake's platform, or get in touch with a member of the Snowflake team.

IN THIS CHAPTER

- Understanding historical precedents
- Evolving from descriptive to prescriptive analytics
- Accommodating all your users
- Utilizing all your data
- Leveraging the power of the cloud

Chapter 1
Defining the Urgency for Modern Cloud Data Analytics

Analytics tools and practices have evolved steadily over the years. But until recently, even cloud data analytics solutions retained the rigid attributes of the on-premises systems from which they were derived. These hasty "rehosting" exercises have prevented organizations from effortlessly scaling their analytic systems and democratizing access to a single copy of their data. To this day, these "cloud-washed" solutions have also made it difficult to easily and securely share data across an organization and beyond — capabilities that can be achieved only when analytic solutions have been natively architected for the cloud.

This chapter reviews the progression of analytics along an upward maturity curve. It explains the technology advancements that have made cloud data analytics so pertinent to today's businesses and reveals how technology leaders can use the cloud to democratize analytics among their entire workforce as they create the types of advanced analytics solutions that allow them to respond to today's constantly changing economic, political, and societal trends.

Tracing Analytics History

With the rise of database management systems, analytics evolved rapidly, thanks to new technologies for storing and processing data. Simple decision support systems of the 1980s advanced to the artificial intelligence (AI)-driven predictive models of today.

In the last ten years, cloud computing paradigms have emerged to extend analytics in exciting new ways. Today, any organization with an urgent need to extract insights from its data can benefit from cloud data analytics due to its inherent speed and scale. Cloud analytic systems can do more than simply provide the infrastructure to store massive amounts of data. Thanks to virtually unlimited compute power, analytic results are virtually instantaneous. This makes cloud data analytics ideal for real-time or near real-time endeavors, such as a marketing team wanting to assess the impact of a limited-time offer or promotion through a social media platform.

The cloud can also facilitate widespread collaboration. For example, during the early months of 2020 when the COVID-19 virus was spreading fast, cloud data platform capabilities made it easy to create a single source for many applicable COVID data sets that spanned organizations and industries. This made data easily available for investigating public health and business impacts, and provided the huge storage and concurrency required.

Evolving from Reactive to Predictive Insights

Traditional business intelligence (BI) tools mainly produce *descriptive analytics,* which are commonly used for historical reporting. These systems allow organizations to measure performance and analyze operational data, such as monthly sales, website traffic, and assessing the results of an advertising campaign. The basic approach is simple: Gather the data and examine it to find out what happened.

Descriptive analytic applications are necessarily backward-looking. For example, a sales dashboard might reveal total revenue over the previous day, week, or month, and broken down by region,

type of product, and other variables. Users can filter the data to select the subsets they're interested in and visualize the results in interactive charts, graphs, and reports. Embedded logic can perform calculations on the data, such as revealing percentages and ranking results. This allows a report or dashboard to transform raw data into meaningful insights, such as which styles of jackets sold best in each store or on the website of a large retailer, and how this week's sales compare to last week's.

Until recently, this type of logic was programmed into BI applications via scripts and queries, often coded in Structured Query Language (SQL). Newer BI platforms include machine learning technology to automatically make these calculations and associations via mathematical algorithms that extract knowledge and insights from a given data set.

Diagnostic analytics delve deeper, with practitioners using data mining tools to spot correlations in the data and querying, filtering, and searching for associations and anomalies to determine *why* an event happened. For example, a clothing retailer might discover that first-time buyers coming from Facebook tend to respond mainly to the specific offers that landed in their news feeds, while return shoppers are more likely to click through to other recommendations, particularly when they have spent time completing an account profile.

Descriptive and diagnostic analytics are a little like looking in the rearview mirror: they make determinations about what happened yesterday or last week by examining historical data. *Predictive analytics* systems help users peer into the future. These systems use machine learning algorithms to determine what may happen tomorrow or next week. They examine historical patterns in the data in conjunction with third-party data sources to assess probable outcomes.

Returning to the shopping example, a predictive analytics system can forecast next summer's sales based on the previous summer's results in conjunction with pertinent economic data, weather forecasts, industry projections, and related factors. Such a system could not only help a buyer for a department store determine how much of each size, style, and brand of jacket to order, but could also automate the replenishment of goods as the stock of these jackets diminishes, to keep the optimum amount of each item on hand.

Driving closer to a real-time replenishment model speeds up inventory turns and minimizes excess stock in warehouses and distribution centers — and can even send demand signals up the supply chain to influence production processes. Instead of spending their time developing manual planning reports, merchandising managers can rely on predictive models that make planning decisions for them, so they can focus instead on vendor negotiations, sourcing, and other items that require human decision-making.

Prescriptive analytics systems go one step further. Based on the results of the predictive algorithms, these applications recommend a specific course of action by considering dynamically shifting variables, such as moment-to-moment sales during a promotion or campaign. These analytic systems know how to maximize the chances of achieving a desired outcome, such as a recommendation engine that can clear out last season's apparel by recommending a new line of complementary shirts, shoes, and pants. With the rise of omnichannel sales, merchants must deliver a connected shopping experience that alerts customers to special offers on products they care about, and predictive analytic systems make it easy for those merchants to identify, purchase, and receive the goods.


TIP

The best prescriptive systems use machine learning (ML) to minimize the need for human analysts. Once data scientists identify the algorithms and train the ML models, the systems make these predictions on their own — and they get smarter over time.

Note that all of these systems have merit, so look for a solution that can deliver capabilities for each (see Figure 1-1).

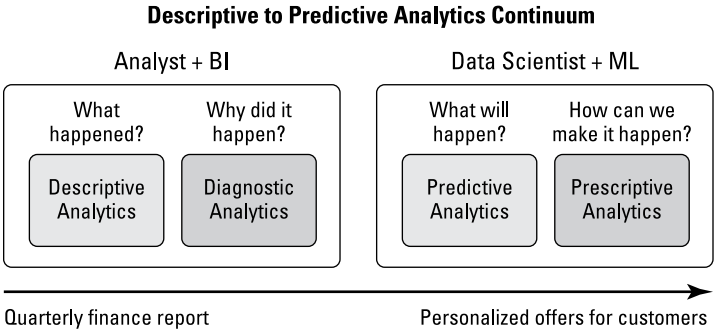


FIGURE 1-1: Understanding the what, why, and how of analytics and using the one that best fits each use case.

Democratizing Analytics

At most organizations, BI and data science initiatives are the province of an elite group of professionally trained analysts. Data is not readily accessible to the majority of the workforce, and proficiency with analytics extends to only highly skilled workers. Democratizing analytics involves breaking down these barriers by extending data-driven decision-making capabilities beyond professional analysts and data scientists. It involves using data to improve every operational aspect of your company by offering analytic capabilities not just to senior managers and executives but also to workers in every department where data can make a difference in their jobs.

Serving all types of users

Success in analytics is directly proportional to the number of people who benefit from your data. Yet not everyone has the same needs and abilities.

Managers and executives prefer self-service apps that make it easy to visualize data via digital charts, graphs, and maps. They want to point and click to obtain the information they need.

Front-line operational workers want information to appear naturally within the context of their day-to-day activities, such as when a call center dashboard automatically displays the details about a particular customer as a call from that customer is received. A few dynamic fields might be all a support rep needs to identify and resolve current issues within that customer account.

Professional analysts need BI platforms and cloud-native tools that allow them to create all of these analytic assets. These analysts are comfortable defining data models, creating analytic portals, and integrating analytics into existing applications and workflows. Data scientists want easy access to data to build and train machine learning models and scoring applications. They want to easily import data into data science notebooks and automated ML tools.

As you assess the potential for analytics within your organization, remember that while nearly every knowledge worker can benefit from analytics, not everybody has the skills or patience to use analytic "tools." You need an analytics strategy that can

accommodate all types of users. The best analytic apps help users at the point of decision: They are naturally infused into the business processes that people depend on to complete their work, and they automatically extend data and decision-making abilities to workers within the context of their work.

TIP

Look for opportunities to "push" information automatically via alerts and screen "pops" and encapsulate analytics into intuitive mobile apps and dashboards. That way, people don't have to go looking for information. It finds them when they need it.

Looking beyond the enterprise

Democratization includes not just internal employees but also external constituents. Extending analytics within and beyond the enterprise is often a two-step process. Step 1 is to *operationalize* insights for decision-making by employees. Step 2 is to share data to streamline operations and look for opportunities to *monetize* data investments by extending analytic capabilities to partners and customers.

CLOUD ANALYTICS DRIVE REVENUE GROWTH

CASE STUDY

As a leading provider of SaaS pharmacy and clinical technology solutions, PDX wanted to help its 10,000 pharmacy customers develop a complete picture of the patients they serve, and then share and monetize analytics as a value-added service. The cloud made it possible.

Previously, limitations with its on-premises database environment made it difficult for PDX to scale its analytic systems to accommodate the millions of daily transactions its customers generate. Its database tables contained nearly 3 billion rows and had to process 5 million new or changed healthcare records on any given day. Properly securing and encrypting all that data to comply with the Health Insurance Portability and Accountability Act (HIPAA) regulations hurt query performance. Facing an expensive hardware and software upgrade, PDX IT leaders decided to acquire a cloud data platform to store and process the organization's growing volume of pharmacy data and extract meaningful insights on behalf of its customers.

Moving analytics activities to the cloud has eliminated many IT tasks related to managing infrastructure, tuning queries, and scaling database management systems. IT administrators can easily create accounts, scale capacity, manage permissions, and produce reports. Security extends from the database up to the application layer, and always-on encryption satisfies HIPAA requirements.

Most importantly, a cloud data platform makes it possible for PDX to monetize its data and create new revenue streams. Authorized users can easily submit queries, mine data, and run analytics against structured and semi-structured data sources, including JSON, Avro, and XML — without taxing PDX's production systems. Analytics and data visualization workloads run simultaneously with data load and ingestion procedures. The near-limitless scalability of the cloud ensures outstanding query performance and eliminates resource contention.

Utilizing all types of data

Having a wide variety of data at your fingertips broadens the scope of your analytic endeavors, from routine financial reporting to advanced data science workloads. A pervasive analytics strategy must accommodate data from spreadsheets, departmental databases, data warehouses, data lakes, and Internet data services, such as real-time weather forecasts and stock market movements, which may take the form of event streams. Today's modern cloud data platforms make this possible.

For example, in addition to accommodating data from traditional data warehouses, which are great for storing relational data in predefined tables, a cloud data platform can ingest and store raw data from weblogs, equipment sensors, social media networks, and other sources that don't conform to a rigid tabular structure. Web data may be stored as JavaScript Object Notation (JSON) files. Spreadsheets may occupy comma-separated value (CSV) formats or tab-delimited text files. And data interchanged among multiple applications may be defined in Extensible Markup Language (XML), complete with tags and other types of artifacts that identify distinct entities within the data.

To take full advantage of the potential of analytics, you need a cloud solution that can easily store, unify, analyze, and share many types of data. The solution should provide convenient *access* to that data, improve the speed at which you can *analyze* the data, and facilitate the process of securely *sharing* it across your organization and within an extended network of customers, suppliers, distributors, and other business partners.

A FOUNDATION FOR INNOVATION

A cloud data platform helps you take advantage of three important technology trends:

- **The rise of the cloud:** Traditional data center infrastructure is sized for a known set of data management tasks. The cloud offers nearly unlimited capacity for storing and processing data, enabling a wide array of concurrent workloads within a centralized platform.
- **The relentless growth of data:** Data will continue to grow in both size and diversity, driven, in part, by the proliferation of Internet, mobile, social, and Internet of Things (IoT) technologies, all of which produce immense quantities of raw data. Within all this new data lie valuable insights for organizations with the technology, resources, and commitment to tap its potential.
- **The importance of analytics:** As the appetite for data continues to grow, analytics will become central to more and more business processes, from traditional management reporting to forward-looking predictive and prescriptive analytics.

IN THIS CHAPTER

- Understanding data services
- Resolving problems with data silos
- Working with data streams
- Reviewing data governance procedures

Chapter 2
Delivering on the Promise of Analytics

This chapter introduces the data services you need to deliver essential and sophisticated analytics, as well as to democratize analytics to a large base of users. It explains the problems that arise from siloed data, reviews the basics of data governance, and describes the advantages of having a centralized source of data maintained in the cloud.

Understanding Data Services

It's easy to become focused on point-and-click tools that can display your data via dynamic charts and graphs: 3D histograms, scatter plots, narrative animations, and digital maps. These visuals can be an effective way for many people in your organization to understand information and make decisions. But those front-end analytic tools must be backed by solid data services.

Data services are required to put your data to work collecting, transforming, delivering, and sharing your data with line-of-business managers, professional analysts, application developers, and data scientists. Ideally, these services converge around a centralized repository or *cloud data platform* that becomes the hub for

consolidating a diverse range of analytic activities. They orchestrate the many activities necessary to store, access, analyze, and share your data with your user community, as well as to populate the analytics applications on which that community depends.

But remember, your long-term challenge is not merely to help users analyze and visualize their data but also to make sure people obtain complete, consistent, and accurate data before they issue queries and generate reports. There's no point in becoming a data-driven organization if the user community doesn't trust your data or the misuse of that data introduces regulatory risks to your organization. A complete analytics practice must address data governance, data quality, data security, metadata management, and a host of other concerns. Without effective access controls, a pervasive metadata layer, rigorous data governance procedures, and other essential data services, you will not be able to confidently mobilize and monetize your data.

REMEMBER

To enable analytics on an enterprise scale, you need a front-end data visualization tool and many back-end data services tied to a centralized data platform that resides in the cloud.

A complete cloud data platform must be versatile enough to accommodate many types of data workloads, including data warehousing, data lakes, data science, data sharing, data application development, and data engineering services that bring data into the platform (see Figure 2-1). The platform should enable secure, governed access to all your data, including structured, semi-structured, and unstructured data.

The Workloads of the Modern Cloud Data Platform

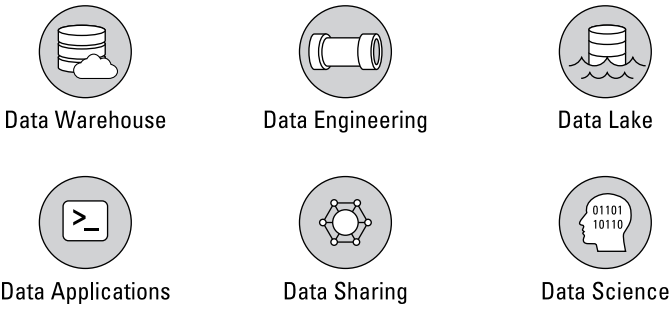


FIGURE 2-1: Establish a central cloud data platform that can accommodate all your data and analytic workloads.

The cloud data platform should be easy to scale as analytic activities increase, and offer unwavering performance as user demands grow. It should be available as a usage-based service that you can easily or automatically turn on and off as needed. And, it should be managed by a cloud vendor dedicated to minimizing administrative chores for your IT staff.

Why does this data platform need to reside in the cloud? The main reason is to maximize your options — to enable your team to access and explore new data sets, pursue unforeseen avenues of inquiry, and find inspiration in unexpected places. There is so much data available, from so many different sources, and in such large quantities that it's hard to manage it all. The cloud offers near-unlimited data storage and computing resources. Cloud implementations are not only less expensive and easier to scale, but they're also virtually trouble-free. All necessary infrastructure and platform services are provisioned as needed, automatically, including the installation of security patches and software updates.


TIP

Look for a cloud data platform that makes it easy to load, store, transform, integrate, analyze, and even monetize near-unlimited amounts of structured, semi-structured, and unstructured data in their native formats.

Resolving Problems with Siloed Data

One of the fundamental principles of this book is for your line-of-business managers, data analysts, data engineers, data scientists, application developers, and frontline workers to all leverage the same single source of data across your organization to ensure consistent outcomes. Their respective jobs are much easier if they can access one repository containing all your data sets and types of data simultaneously without having to import or export data from one system to another. Having a single source of truth also accelerates *time to insight* because users spend less time wrangling data. This benefit is especially important to data engineers and data scientists, whose time is too expensive to waste.

In practice, rallying the enterprise around a single source of truth is rarely this seamless, mainly because of how corporate information systems have been designed and implemented over the last several decades. On-premises or in the cloud, each production application creates its own data silo, with marketing data

residing in a marketing automation system, sales data residing in a customer relationship management (CRM) system, finance data in an enterprise resource planning (ERP) system, inventory data in a warehouse management system, and so on.

These disparities create a domino effect in the analytic databases derived from these production systems: data warehouses for operational reporting, data marts for departmental analytics, and data lakes for data mining and data exploration. Each of these may depend on specialized extract, transform, and load (ETL) tools to collect data from the production systems and transform that data for analysis. The situation has become even more complex with the rise of thousands of software-as-a-service (SaaS) and mobile apps that handle discrete tasks, each with its own source of data.

Engineering data pipelines

Modern data pipelines help resolve these disparities by orchestrating the exchange of data among many different databases and computing platforms, rationalizing the differences among data types, and loading data from many sources into a common repository. Creating a single source for all types of data and all types of workloads will simplify your infrastructure.

Accommodating external tables

Even when most of your data is maintained in a centralized repository, it's still possible to accommodate data in *external tables* (read-only tables that can be used for query and join operations) and *materialized views* (database objects that contain the precomputed results of a query). This versatile architecture enables seamless, high-performance analytics and governance, even when the data arises from more than one location.

Once your data silos are rationalized into a common source in this way and acted on by a common set of data services, it's much easier to maintain and govern that data, keeping it clean and consistent. This paves the way for self-service analytics, which is covered in Chapter 3.

Dealing with streaming data

In addition to accessing and maintaining data from databases, your analytic systems may need to load and process data from event streams. Popular streaming data sources include weblog

data from Internet viewing and browsing activities, as well as continuous sources of Internet of Things (IoT) data emitted by sensors from factory production processes, supply chains, transportation networks, and many other sources. Most commonly, this data can be ingested and updated in "near real-time," such as every five minutes. Depending on the application, however, some streaming data must be loaded and processed continuously.

Fraud detection services operate according to the principles of streaming data, as do recommendation engines used by entertainment companies such as Netflix, which constantly monitor viewer activity to determine what is popular and make real-time recommendations as soon as you turn on your TV.

Most event streams comprise *time series data* — a sequence of data points indexed temporally. Examples include tide levels measured by a weather buoy, a patient's heart rate transmitted via a cardiac monitor, transaction logs, or the New York Stock Exchange's daily closing value. Time series analysis involves analyzing these data streams to extract meaningful statistics, such as forecasting future values based on previously observed values.

Introducing Data Governance

REMEMBER

A good analytics practice stems from good data governance practices: a pervasive strategy that seeks to maintain high-quality data throughout its lifecycle, with consistent controls to support the organization's business objectives and meet regional and industry data regulations.

At its most basic level, *data governance* entails knowing precisely what data you have, where it resides, who is authorized to access the data, and how those authorized are permitted to use it. These activities are extremely important. However, in a report titled "Data-Forward Enterprise: How to Maximize Data Leverage for Better Business Outcomes," IDC reported that nearly 46 percent of organizations struggle with data governance, citing data management deficiencies.

To minimize this struggle, attend to the basics: data availability, usability, consistency, integrity, and security. The objective is to reduce the risk of compliance violations, protect sensitive data, and minimize the adverse effects of poor data quality as data is

disseminated and shared across the organization. It's important to have procedures in place that prevent the unauthorized transmission of sensitive data. Incorrectly exporting, copying, and combining data causes data silos, which complicates data governance and compromises data security.

Controlling data access

Cloud data governance starts with knowing where your data comes from, where it resides, who has access to it, how it's used, and how to delete all instances of it when you are required to do so.

Different types of users have different needs. By adopting role-based procedures, you can assign specific levels of access for each type of user. This approach ensures users can access only the data they're permitted to see. You can also employ data masking to limit visibility to your data at the column level (such as masking salary data or Social Security numbers).

These procedures control data on a "need to know" basis. For example, a line-of-business manager needs to see the monetary values within a salary report. But the database administrators who maintain the application don't need to see employee Social Security numbers. They simply need to know they are properly displayed within a nine-character field.

Examining security basics

Cloud data security in this context also starts with knowing where your data comes from, where it resides, and who has access to it.

All your analytic applications should consistently authorize users, authenticate their credentials, and grant access only to the data they are authorized to access. Work toward centralizing your data sources, along with the associated user authentication.

For sensitive business data, consider adding multifactor authentication for a second level of identity verification by sending temporary security codes to a user's email address or mobile phone number. You'll also want to employ basic database security to control access to the data at multiple levels, including all database tables and schemas. Database security procedures determine who can access the database objects and what operations they can perform.

IMPORTANT ASPECTS OF DATA SECURITY

A complete data security strategy protects your data in transit and at rest.

Access control: Ensures users can access only the data they're permitted to see. Access control should be applied to all database objects, including tables, schemas, and any external tables.

Data protection, retention, and redundancy: In case of a mishap or a malicious attack, you should be able to instantly restore or query previous versions of your data in a table or database within a specified retention period.

Data encryption: Data should be encrypted both when it is "at rest" (stored) and in transit (moving to and from the database for display to a user or application). Query results should also be encrypted.

Additional security methods include full or partial data redaction and end-to-end encryption that keeps data illegible until it is accessed by an authorized user or displayed in a sanctioned application.

Enforcing consistency with metadata

Pulling all your data into a centralized repository allows you to establish a consistent metadata layer. Metadata describes what data is used for. It makes tracking and working with data easier. It ensures all users obtain consistent results, and all workloads deliver consistent outcomes, no matter how many queries and transactions are conducted. Metadata is the foundation to ensuring users have an easy time finding, viewing, tracing, and organizing data for analysis.

Paying attention to data quality

All analytic applications must be fed by clean, accurate data and governed by consistent data quality procedures. A complete data governance strategy includes overseeing the quality of data coming into an organization and ensuring the consistent use of that data as it is shared throughout the organization and beyond. You

must be able to identify when data is corrupt, inaccurate, out of date, or incomplete. However, a recent Dataversity report titled "The 2020 State of Data Governance and Automation" revealed that 58 percent of companies have difficulty understanding the quality of their key data sources. Many of the IT professionals IDC surveyed admitted that ensuring good data quality was one of the most serious issues in their analytics practice.


TIP

To rein in problems with data quality, you need to enlist help from the people who are closest to the data. These *data stewards,* generally drawn from the business rather than IT, set data quality rules and monitor data quality processes based on their understanding of the functional domain, whether finance, sales, manufacturing, customer support, or any other corporate function. Because these individuals are closest to the data, they are the right candidates to identify when that data is corrupt, inaccurate, or not refreshed often enough to be relevant.

Cataloging data in a schema or catalog

One way to track the lineage of your data and monitor who can access that data is to create an information schema, also known as a *data catalog*. This gives you visibility into the metadata to see how data objects are being accessed, changed, and moved. A data catalog tool lets you connect to cloud and on-premises data sources. It assists with data profiling, cataloging (including lineage), and integrating data into your central data repository, where it enables IT and business teams to collaborate, share, and access trusted, governed data from a central location.

Complying with government and industry regulations

This chapter provides general guidelines regarding instituting governance procedures to control the ownership, use, and accessibility of your data. Some of these practices become specific mandates in the context of government and industry regulations, such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI DSS), and the Sarbanes-Oxley Act (SOX). The types of information that commonly fall under these guidelines include credit card information, Social Security numbers, dates of birth, IP network information, and geolocation coordinates.

5 STEPS TO GOOD GOVERNANCE

Implementing effective governance at the outset will help you avoid potential pitfalls later on, such as poor access controls, inconsistent metadata management, unacceptable data quality, and insufficient data security. To get started on the right foot with enterprise analytics, follow these steps:

1. **Centralize your data:** Siloed information makes it difficult to trace the data's lineage, catalog the data, and apply security rules. Combining your data into a centralized repository simplifies these tasks.
2. **Employ rigorous data quality procedures:** Data quality allows organizations to trust their data and make good decisions, even when the data is derived from many different sources. Enlist *data stewards* to set data quality rules and monitor data quality processes based on their understanding of each functional domain.
3. **Create a data catalog to define all data:** What data do you have, where is it, and who can access it? Track the lineage of your data and govern access to that data via a centralized data schema or dictionary.
4. **Enforce consistent data security procedures:** Having a centrally governed repository of data, maintained in the cloud, ensures good governance and secure outcomes. In addition, you should encrypt your data, adopt role-based security procedures, and employ database security. Consider using data masking, data redaction, and multifactor authentication for sensitive data.
5. **Put regulatory compliance on autopilot:** Set up rules to control data ownership, data access, data usage, and auditing, with special attention to data bound by industry regulations.

If you already have a comprehensive data governance strategy, these regulations will be easy to adhere to. For example, they typically require organizations to trace their data from source to retirement, identify who has access to it, and verify how and where data is used, with complete audit capabilities.

REMEMBER

Poor data governance can result in financial penalties, lawsuits, and even jail time. To minimize these risks, eliminate disparate sources of data and enforce consistent administrative practices. Having complete transparency into the source and lineage of your

data makes it easier to audit results and generate compliance reports. Advanced governance technologies, such as dynamic data masking and secure views, give you additional options for protecting sensitive data.

Adopting Advanced Analytics

Having a consolidated source of clean, accurate data enables you to take advantage of more sophisticated use cases, such as machine learning (ML). If you plan to go down this road, you'll want a data platform that allows you to amass all your data for ML initiatives and supports read/write integration with automated machine learning (AutoML) tools. For example, bidirectional integration between a cloud data platform and an AutoML tool enables users to store and query data in the cloud platform, create predictions in the AutoML environment, and send results back to the platform, removing the need for complex programming. Such integrations open up machine learning projects to users of any skill level. But remember: Your machine learning algorithms are only as good as your data.

Embedding Analytics into Applications

Keep in mind that until employees on the front lines of your business can obtain trusted data at an operational level, your company will face a huge barrier to its potential success. One way to broadly extend cloud data analytics throughout the enterprise is to embed simple search, query, and data visualization capabilities into the applications employees already use. Embedding analytics involves adding discrete functions normally associated with business intelligence (BI) software, such as dashboard reporting, data visualization, and analytics tools, to existing applications.

You'll want analytics tools that make it easy to create dashboards, define KPIs, and embed analytic functions into a web or mobile app. For example, a manufacturer might create an embedded analytic environment that allows transportation managers to select the best carriers for each shipment, monitor deliveries, and generate freight audit reports. A grocery retailer might want

simple BI capabilities to monitor the success of daily merchandising promotions, with links to the warehouse management system to adjust stock levels based on daily consumer demand. Chapter 3 explores these and other analytic scenarios in detail and discusses the technology solutions that make them possible.

The Value of a Cloud Data Platform

Remember, to maximize the potential of your analytics endeavors you need a scalable cloud data platform that can bring together the optimal set of resources for each analytic scenario (see Figure 2-2). Chapter 3 explores these scenarios in greater detail and discusses the cloud data platforms that make them possible.

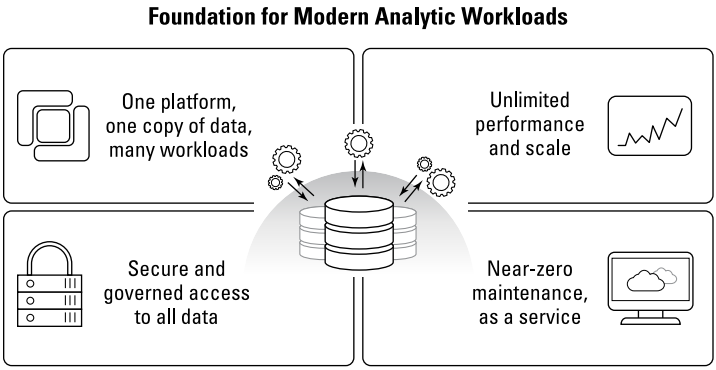


FIGURE 2-2: A cloud data platform offers the data and processing foundation for modern analytic workloads.

CONSOLIDATING DISPARATE DATA INTO A SINGLE SOURCE OF TRUTH

CASE STUDY

Coupa Software offers a cloud-based business spend management (BSM) software platform, which connects hundreds of Coupa customers in the Americas, EMEA, and APAC with millions of their suppliers globally.

(continued)

(continued)

Over a two-year period, Coupa grew from 700 employees in five locations to 1,500 employees in more than 20 locations and simultaneously acquired and integrated four companies. In the wake of this sudden growth and diversification, Coupa's Business Systems team had to integrate a variety of disparate systems, processes, and data. Lack of data governance created issues with data integrity, deprecation, and regulatory compliance. As a publicly traded company, Coupa must comply with federal requirements, such as the Sarbanes-Oxley Act, and international regulations, such as GDPR.

Coupa established a single source of truth for business data across its rapidly expanding enterprise by standardizing on a cloud data platform known for its strong data governance capabilities. Today, the business systems team can easily support the internal business intelligence needs of sales, finance, customer support, customer success, operations, and marketing.

The platform anchors a comprehensive data strategy that makes it easy to comply with regulatory requirements. The team can focus on gathering business insights rather than managing hardware and software infrastructure. A flexible billing model ensures Coupa pays only for the compute resources it consumes, while on-demand scaling aligns usage costs to business growth. Going forward, Coupa plans to create additional analytic assets to support product management, predictive sales models, and customer behavior.

IN THIS CHAPTER

- Enabling self-service with governance
- Creating business intelligence assets
- Moving beyond historical analysis
- Simplifying data science workflows
- Writing effective queries
- Standardizing on SQL

Chapter 3
Unlocking the Power of Your Data

This chapter describes how to use a cloud data platform for your analytics data foundation. It lays out the fundamental tenets of business intelligence, data science, good query performance, and using Structured Query Language (SQL). Specifically, it introduces basic data science and machine learning concepts and explains the importance of SQL as a standard language of data.

Building Organizational Trust in Cloud Data Analytics

In the past, a few back-office workers controlled business analytics, and data management was much simpler. Only a small number of people were allowed to touch the data. They were mostly professionally trained analysts who understood the need for constraints and controls. They knew that data mining, data exploration, and ad hoc data analysis could place a drain on IT resources and even expose their organizations to security and compliance risks.

However, the premise of modern analytics is to extend data-driven decision-making from a few dozen highly skilled people to hundreds or even thousands of workers throughout your organization. To roll out self-service analytic capabilities to your entire organization, follow the rigorous data governance recommendations discussed in Chapter 2.

REMEMBER

These data governance principles are so important they are worth repeating. Make sure that your data is clean, consistent, and up to date. Verify all users understand the data's intended purpose and lineage. Don't store duplicative copies of data. Confirm everybody accesses the same data in the same way: For example, if managers in two different departments calculate a profit-and-loss ratio, their results will be the same. Too often, that is not the case.

Finally, make sure your cloud data platform is architected to accommodate high concurrency so escalating user activity doesn't bog down your IT infrastructure or lead to impatience among users if queries don't return results quickly. Separating storage from compute will help you avoid the proverbial Wild West of self-service analytics: nobody marshaling the data and everybody taking matters into their own hands, with no centralized control.

Laying a Foundation for Pervasive Business Intelligence

Business intelligence (BI) refers to a broad class of technologies and tools that help business users analyze data for historical analysis. BI enhances business processes by turning data into useful information and builds knowledge by encouraging inquiry and exploration into various business domains, from production to sales, finance to HR, and manufacturing to transportation.

BI solutions are used to produce business and financial reports, conduct self-service data exploration, and provide the business with packaged dashboards, portals, scorecards, and many types of "front-end" assets. These solutions also enable advanced capabilities, such as geospatial analytics and graph analytics.

TIP

Some BI solutions include user-friendly visualizations that help users of all skill levels explore data without writing queries. These solutions also allow people to quickly develop dashboards and share them with other members of the organization.

Rich visualizations and out-of-the-box widgets help users move quickly from data to insights.

Discerning the difference between facts and intuition

Analysts use a combination of tools and intuition to define hypotheses and then determine whether the data confirms or contradicts each hypothesis. They use BI solutions and data visualization tools to discover trends and patterns in data, and then use their domain knowledge or business processes to explain why those trends and patterns occurred.

For example, a grocery supply chain manager might notice that fresh-baked goods sell best on Thursday and Friday, especially in stores located in popular vacation destinations, and thus decide to increase the production of certain items to accommodate anticipated demand. The data revealed observable sales trends, but the decision to boost the supply of baked goods was based on an intuition that these trends would continue.

Moving beyond historical analysis

To bolster intuition with factual insights, many companies have started to shift from basic reporting and historical analysis to using advanced mathematical models and machine learning (ML) algorithms to find patterns that can predict future outcomes. Often, the analytics progress from descriptive to predictive to prescriptive analytics, as described in Chapter 1.

For example, an ML model will not only observe and report on trends in the sale of baked goods but also predict which types of baked goods will sell best in each store and region on certain days and times. It might also suggest how to optimize staffing levels in the bakery based on these trends, or even suggest a new bakery location to minimize transit times to ensure baked items reach the display cases while still fresh. The machine reads the data and makes the recommendations.

Technically advanced companies have used these types of predictive and prescriptive analytics successfully for years. For example, Netflix knows how to increase the viewership of a new show by recommending it to viewers who like similar shows. Uber can direct its drivers to certain areas in anticipation of demand, such as when a scheduled concert or event lets out, and adjust pricing

as requests increase. Its ML models are constantly analyzing driver and passenger data in conjunction with data about current events, weather, and many other variables to automatically adjust rates and staffing levels.

REMEMBER

ML models make these advanced analytic scenarios possible. There's so much data that human brains can no longer comprehend all of the variables. We can't recognize the patterns, and we can't discern, let alone remember, all the trends. It is quickly becoming a job for data scientists and the algorithms they create to automate these decision-making processes.

Understanding the Potential of Data Science

Data science includes tools and techniques for analyzing large amounts of data, including the ML models that process it. Just a few years ago, a data science practice was a rarity, the province of big companies with large IT budgets. Today, data science and ML have entered the mainstream. In a 2019 TDWI Best Practices Report titled "Driving Digital Transformation Using AI and Machine Learning," 92 percent of survey respondents reported using ML technology, and 85 percent said they are building predictive models using ML tools.

The need for data scientists is growing at a rapid pace as these analytic activities impact many business functions and automate common tasks. Many common use cases have emerged. For example, marketing analytics help companies attract and retain customers based on insights derived from social media data, email, and clickstream data. ML models help segment customers, predict churn, improve retention, and recommend products based on buyer behavior.

Fraud and risk models help insurance companies detect suspicious claims and help finance companies identify excessive credit risks. For example, an auto insurance company might collect Internet of Things (IoT) data from vehicles to monitor driving habits, such as miles driven, smoothness of acceleration, speed, and braking habits, to offer discounts to careful drivers. A healthcare company might create an ML model to examine the likelihood of patients developing infections during hospital stays based on the time

of year, hospital load, and mitigating health factors. Healthcare providers can study population health data to determine how the interaction of various pharmaceutical compounds affects health outcomes. A university can use data science to evaluate the risk of students dropping out of school, and factories can analyze environmental and maintenance data to predict and provide preemptive maintenance so assembly lines don't break down. Shipping companies can study transit times and optimize transportation routes based on real-time weather and traffic data.

All of these scenarios have one thing in common: a willingness to use both public and private data to predict upcoming needs and forecast likely outcomes.

REMEMBER

Data scientists need easy access to complete and accurate data and powerful tools for exploring and understanding it. This includes dedicated compute resources that can streamline data preparation, an easy way to connect to ML tools, and the ability to incorporate the output of ML models into business processes and applications, often via more approachable BI tools that are accessing this same data.

Improving data science workflows

Data scientists require massive amounts of data to build and train the ML models that power these predictive use cases. Unfortunately, many of these workers complain that they spend the majority of their time cleaning and preparing the raw data rather than building and training models. There are six basic steps in this data-intensive process (see Figure 3-1):

1. **Collecting data:** For each business problem data scientists try to solve, they must identify pertinent data sets, both internally and from external third parties.
2. **Exploring the data:** Next, data scientists try to understand the data, typically through visualization, to discover patterns and attributes that might indicate patterns.
3. **Feature engineering and transformation:** Once they have identified pertinent data and understand its potential, data scientists need to cleanse it, transform it, and perform *feature engineering* — the process of identifying common attributes from existing data and creating new features to improve the performance of ML algorithms.

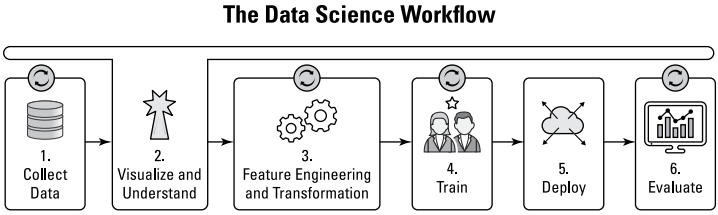


FIGURE 3-1: The data science workflow for populating, training, and deploying ML models. Visualizing results can span the entire workflow.

4. **Training the model:** After the data has been prepared and new features engineered, scientists use it to "train" ML models to identify patterns. These models must be retrained periodically, which requires fresh training data sets to be prepared via the same cycle.
5. **Deploying the model:** Once the model is trained, it is deployed via an API, embedded into an application, or incorporated into BI reports so the business community can begin deriving insights.
6. **Evaluating outcomes:** Each model is monitored and evaluated based on how its predictions compare to actual outcomes over time.

Consolidating data for BI and ML in a central location is the starting point for streamlining this data-intensive workflow. It also makes it easier to synthesize and use this data in a wide range of data science notebooks and ML tools (see Figure 3-2). Having a common repository allows business intelligence apps to leverage the results of data science initiatives and put the data to work. This unified approach allows data scientists to output the results of ML activities back into the data platform for general-purpose analytics. All front-end tools reference the same back-end data definitions, ensuring consistent results for queries, forecasts, dashboards, and reports.

TIP

Having a central repository also streamlines access and simplifies permissions so users don't need to wait to receive permission to access new data. Some data platforms include an extensive ecosystem of technology providers that have integrated their data science and ML tools, simplifying activities for a data science team.

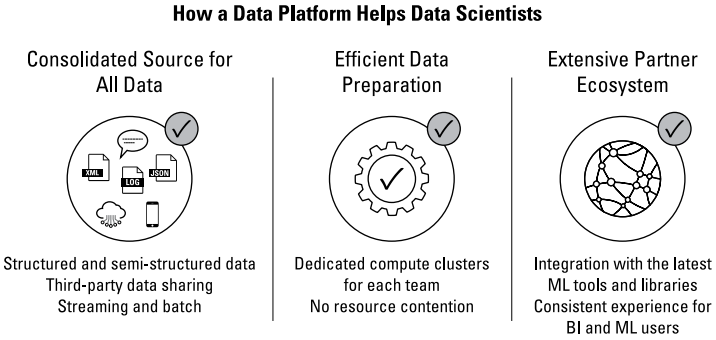


FIGURE 3-2: A cloud data platform enables fast, centralized access to machine learning data, provides dedicated compute resources for processing and preparing that data, and enlists potential help from a large partner ecosystem.

Automating machine learning tasks

Scala, R, Java, Python, and other popular languages empower data scientists to develop and create predictive analytics applications and train the associated ML models. New AutoML tools streamline this data science process by automating much of the manual coding work. For example, instead of requiring analysts to crunch the data and develop hypotheses, they can depend on an AutoML tool to examine the data, recognize correlations, and make automated suggestions.

AutoML tools select and simultaneously train dozens of possible ML models and apply the most accurate one to the analytic problems at hand. They recognize the type of data each model requires, such as time-series data, retail data, or customer data, and suggest the best algorithms based on their understanding of how these types of data are normally used. These tools also know what types of input they need, whether that be cardinal data (counting numbers) or Boolean data (data with two possible values, such as true and false) or nested text fields (semi-structured data), and they can prepare that data automatically. Previously, highly trained data scientists, always in short supply, had to spend time manually figuring all this out.

AUSTRALIAN UNIVERSITY STREAMLINES ANALYTICS

CASE STUDY

Founded in 1850, the University of Sydney is one of the world's leading universities. The institution's 8,000 staff members deliver education support and services for more than 60,000 students.

During 2018, the university experienced performance and capacity issues with its on-premises data warehouse. Ingesting, managing, and querying large data sets was difficult. As the volume of data grew, it became nearly impossible to produce key reports, some of which involved combining multiple data sets to deliver needed insights.

After examining several on-premises and cloud-based options, the university subscribed to a cloud data platform to extend analytic capabilities throughout campus. Thanks to the unique cloud-native architecture of this system, which separates storage and compute resources into independently scalable entities, the university's data analysis and report generation capabilities instantly improved. Analysts can easily query data sets with more than a million records, and sophisticated queries that used to take hours to complete now run in minutes. Because the university pays only for the capacity and resources it uses, operational costs have dropped by 25 percent.

The University of Sydney's new BI platform has provided invaluable support during the COVID-19 pandemic closures. The analytics team combined student data with publicly available data sets to navigate the shutdown and prepare for the resumption of campus activity.

The university plans to make analytics available to additional educational and administrative groups based on a single source of truth derived from a variety of data sources. As the university continues to democratize analytics, its versatile cloud data platform can handle an escalating volume of data, users, and workloads.

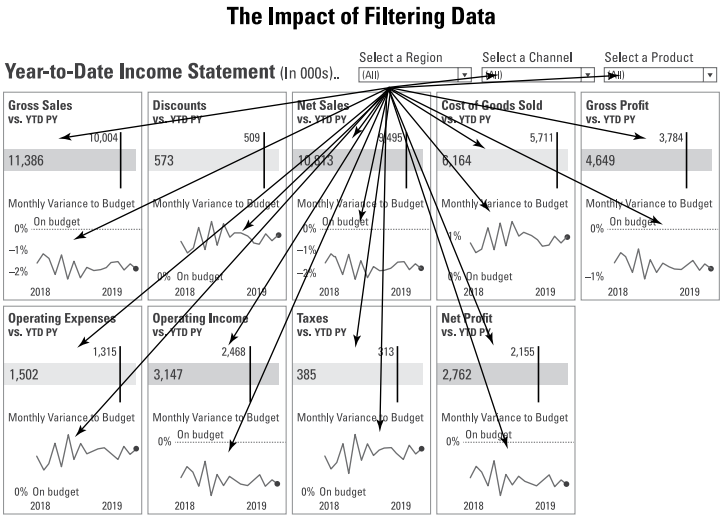
Writing Effective Queries

As organizations extend analytics more broadly, increased query activities can strain the underlying data infrastructure. For example, in addition to traditional reporting, where queries run at scheduled intervals to populate a known set of reports, as you democratize BI, an increasing number of users will start querying the database in a "self-service" fashion, using drag-and-drop tools.

TIP

Using BI tools is relatively easy, which may mask the complexity of the underlying data operations. In some cases, one simple operation can fire off dozens of queries. Every time you manipulate a data object within one of these visual environments, such as refreshing a chart or drilling into a map, you are initiating query operations to sort, summarize, filter, and otherwise manipulate what might be millions of rows of data.

For example, Figure 3-3 shows a finance dashboard that helps managers monitor year-to-date income as they filter revenue data by region, channel, and product. When a user enters data into the three fields at the top of the screen to refresh the data in the dashboard, all nine analytic zones must be redrawn, as shown in the figure.



Source: Tableau Public

FIGURE 3-3: The impact of refreshing a finance dashboard: multiple simultaneous queries against the database.

One way to improve query performance is to restrict the answer set so only the first 10 or 20 rows of data are retrieved. As a rule of thumb, you should only return as much data as a human can reasonably comprehend at one time. This method makes it easier for the user and helps prevent "runaway queries" that consume excessive computing resources.

TIP

Another effective technique is to push subsets of data or summaries "up the stack" so that multiple reports, dashboards, and other front-end applications can utilize them. For example, the logic for totaling revenue by quarter can be written as a stored procedure in the database schema and the query results cached for repeat access. That's especially important when multiple users are accessing the same dashboard. Whenever a user refreshes the dashboard view, the results should be available instantly, without the system having to run the calculations all over again.

As mentioned in Chapter 2, you can also use materialized views to precalculate metadata and statistics about data in external files, thus speeding up the queries that are run against the database. Materialized views store the results of SQL SELECT statements and maintain the view of data in a form multiple applications can access. When the underlying table changes, the materialized view is automatically updated. The idea is to *materialize* the data you query most frequently. Some BI tools use the term "data extracts" to describe these views because they only have to be run once, and then all subsequent queries can run against the extract.

Materialized views are automatically refreshed in the background. No matter which BI or analytics tool needs the query results, they will be computed and presented consistently and made broadly available. Users can access the results without having to run the calculations and manipulate the data each time.

Leveraging SQL for Multiple Activities

A diverse team of analysts and data scientists can share a universal source of governed data without having to copy data or move it from place to place if your cloud data platform includes a SQL abstraction layer that can query structured, semi-structured, and unstructured data. With SQL, you can load raw semi-structured data in its native format, such as Extensible Markup Language (XML) or JavaScript Object Notation (JSON), and query it directly with SQL statements without preprocessing.

SQL can be a boon for data scientists, mainly because it's much more efficient and easier to use than Spark, Scala, and other specialized languages. You can even conduct quick and easy feature engineering using SQL, often accomplishing in 100 lines of SQL

what would take thousands of lines of code in Scala. Additionally, SQL procedures execute much more efficiently, and sometimes as much as ten times faster than Spark procedures. Using SQL is also less expensive because you don't need specialized Spark clusters but can run SQL workloads on commodity cloud infrastructure.

TECHNICAL STUFF

Data engineers can use SQL to power data ingestion processes, either in batch mode or via a live stream. Some data platforms use *streams* and *tasks* to maximize the performance of these operations. A *stream* is a special object type that uses change data capture (CDC) technology to track the ongoing changes in a table, including inserts and data manipulation language (DML) changes. Streams allow BI and data science apps to query just the incremental changes rather than the entire data set. A *task* defines a recurring schedule for executing SQL statements, including statements that call stored procedures. You can chain tasks together to support complex processing scenarios. In a continuous data pipeline, tasks can use streams to process new or changed data.

Thanks to the virtually boundless resources of the cloud, these data ingestion workloads don't have to compete with data science queries for the same resources. Each workload can have dedicated compute resources, which can be independently scaled yet still access a single consistent data repository. Independent scalability (the ability to separate compute from storage and compute from other compute) is a fundamental tenet of a modern cloud data platform. Chapter 4 examines this and other architectural principles as it explains how to integrate cloud data analytics into your technology ecosystem.

CASE STUDY

DATA-DRIVEN GROCERY DELIVERY

Good Eggs is an online grocery delivery service that provides fresh local produce, meal kits, grocery staples, household products, and wine to customers throughout the San Francisco Bay Area. Advanced analytics are key to its business model, but Good Eggs' previous data solution was not fast or flexible enough to handle an evolving set of BI and data science initiatives. Good Eggs decided to move its data and analytic workloads to the cloud.

(continued)

(continued)

Today, structured business data from Good Eggs' warehouse management system is loaded alongside its semi-structured customer data into a modern cloud data platform. A special VARIANT data type allows analysts to load semi-structured data in its native form and query it in the same way that they query structured data — via SQL. This versatile cloud architecture allows several groups of users to analyze, visualize, and share data and insights. For example, the warehouse operations team uses the cloud data analytics environment to track the status of delivery routes. The business team uses it to generate reports on average order value and gross margin. The data science team uses a collaborative data science platform to leverage the same data as it builds predictive models to forecast business trends. Each team can spin up its own compute resources and allocate dynamic amounts of computing to queries based on fluctuating needs.

In addition to enabling pervasive analytics with minimal administration, the cloud data platform has alleviated the performance bottlenecks that plagued the previous data solution. In the future, Good Eggs plans to set up automated daily data loads, using a continuous data ingestion service and utilizing the inherent data sharing capabilities of the cloud data platform to securely exchange governed data with its business partners.

IN THIS CHAPTER

- Establishing the right architecture
- Utilizing multiple clouds
- Going deeper with visualization
- Automating integration tasks
- Standardizing on SQL

Chapter 4
Integrating Cloud Data Analytics into Your Technology Ecosystem

To analyze your data, you must be able to store it, connect to it, and present it in meaningful ways, on an enterprise scale. This chapter describes the technology ecosystem you need to fulfill this vision, with an emphasis on the type of *architecture* you need to empower a near-unlimited community of users to intuitively visualize and share data. It explains how to integrate many sources of data into your cloud data repository via flexible data pipelines, and how to establish one unified copy of data to support many different workloads — with instant elasticity, minimal cost, and exceptional performance.

Understanding the Limitations of Traditional Architectures

As you concentrate on the technologies that enable your user community to be productive, don't lose sight of the end goal: to help knowledge workers make fact-based decisions. Cloud data

analytics is the glue that connects those business users with an ever-expanding universe of data resources and sources.

Unfortunately, many of today's popular data platforms were architected as offshoots of legacy, on-premises environments. They were designed to work on closed networks with local data, and they work fine when only a handful of analysts use these analytic environments. However, as organizations deploy larger and more diverse data workloads, legacy platforms slow down under the weight of too many concurrent users and too much data. The infrastructure gets in the way and ultimately distracts from the goal by hindering analytic activities and placing excessive demands on the IT department.

Many of these demands center on preparing the data. Consider the traditional architecture that has characterized business intelligence (BI) initiatives for the last 20 or 30 years. IT professionals orchestrate the movement of data through several phases of integration, transformation, and normalization before analytics are accessible to consumers, as shown in Figure 4-1.

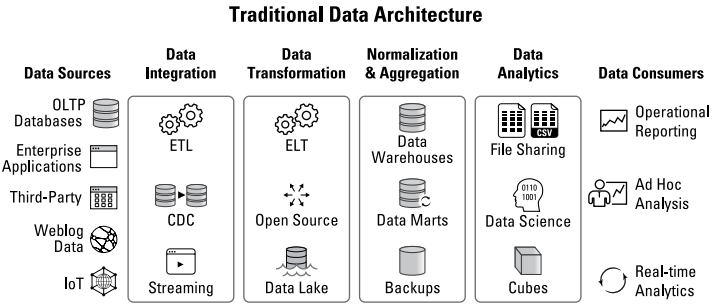


FIGURE 4-1: An intensive approach to gathering and preparing data for analysis.

This approach presents many problems:

- **Complexity:** Each data source requires individual management, security, and infrastructure.
- **Fragmentation:** Having disparate data platforms creates silos, with the consequent problems described in Chapter 2.
- **Limited sharing abilities:** Data must be moved or copied to be shared, restricting collaboration and compromising security and governance.

- **Lack of scalability:** Without the elasticity of the cloud, traditional on-premises architectures can't easily flex to match ever-expanding business needs.
- **Excessive costs:** Data storage and compute resources must be sized to match peak loads, forcing organizations to purchase more resources than they need (and which they will eventually exceed).

Outlining the Advantages of a Cloud Data Platform

The right data platform, built for the cloud, ties together all the components of the technology stack to produce positive *experiences* with your data. It lets you roll out whatever data applications you need without the costly and disruptive impact of having to add more infrastructure to a data center or manage such a complex hardware/software environment. Most importantly, it works with every part of your technology ecosystem as the central hub that streamlines data integration, BI, and advanced analytics.

Three primary motivations exist for using a cloud data platform: to minimize IT maintenance, to reduce overall costs, and to achieve exceptional performance.

Minimizing maintenance

Whether you use traditional BI tools such as Cognos and MicroStrategy, newer BI tools such as Tableau and Looker, or advanced analytic technologies such as DataRobot and Zepl, the goal is the same: to establish a high-performance, zero-maintenance analytics environment with no hardware or software to manage, virtually no queries to tune, no backups to complete, and no security patches to implement.


TIP

A cloud data platform must simplify IT tasks by relieving technology professionals from the burden of overseeing on-premises hardware and software. Within the context of analytics, a cloud data platform streamlines the entire lifecycle of how data is captured, maintained, and used. A cohesive layer of services automates everything from how data is stored and processed to

how transactions are managed, data is secured, and metadata is extended to analysts. Your analytic teams can spend time working with data, not managing infrastructure.

Mitigating costs

Rather than over-provisioning capacity to meet occasional peak demands, your cloud data platform should offer per-second billing to enable each user and workgroup to pay only for the storage and compute resources they need, so the organization as a whole never has to pay for idle capacity. Some cloud vendors charge a flat rate for storage and require a minimum subscription amount no matter how much you use the system, so look closely at prospective vendors' terms. It's far more economical to be able to pay only for actual usage. Furthermore, you should be able to scale compute and storage resources independently under this usage-based pricing model. The vendor should bill you only for the resources you use, and you should be able to suspend compute resources when you stop using them.

Maximizing performance

In a global organization, a rapidly expanding BI practice can place crushing demands on compute and storage infrastructure. As illustrated in Chapter 3, database servers sized for finite workloads can get bogged down by a single user issuing a runaway query. Modern cloud data platforms employ a *multi-cluster, shared data architecture* in which each workload uses dedicated computing resources that can scale independently and elastically. This makes it easy to add capacity as you leverage the near-infinite resources of the cloud.

For example, if you have deployed customer relationship management (CRM) analytics to a large, diverse sales organization, the entire sales team can query the CRM data without degrading performance. As sales activities intensify toward the end of the month or quarter, additional compute resources can be provisioned automatically. You can even run large, resource-intensive data-ingestion workloads at the same time as analytics activities intensify, such as streaming data through an engineering pipeline or training a machine learning (ML) model.

A SCALABLE, HIGH-PERFORMANCE ARCHITECTURE

A multi-cluster, shared data architecture includes three layers that are logically integrated yet scale independently from one another:

- **Storage:** A single place for a single copy of all data
- **Compute:** Independent compute resources dedicated to each individual workload to eliminate resource contention
- **Services:** A common services layer that handles infrastructure, security, metadata, query optimization, and many other integral functions

REMEMBER

Insist on a cloud data platform that independently scales storage and compute as well as compute from other compute resources. This makes it easier to meet the shifting demands of both analytic workloads and data-ingestion workloads, as well as to accommodate peak usage periods. It also makes it easier to gain more favorable pricing terms.

Utilizing Multiple Clouds

A well-architected cloud data platform dynamically brings together the optimal set of resources for each analytic scenario, with a unified code base that spans popular cloud object stores such as Amazon S3, Microsoft Azure Blob Storage, and Google Cloud Storage. This cohesive layer of services maximizes your technology stack options because each public cloud provider has different strengths in different regions, and each cloud service addresses slightly different needs (see Figure 4-2).

Rather than forcing all business units to use the same cloud provider, each business unit can use the cloud services that work best for its unique needs. The best data platforms have a cloud abstraction layer that ensures users will have a cohesive experience across all of them. These "cross-cloud" data platforms

make it easy to share data, ensure business continuity, and comply with industry-specific and region-specific data sovereignty requirements — without worrying about being "locked into" a particular vendor's cloud.

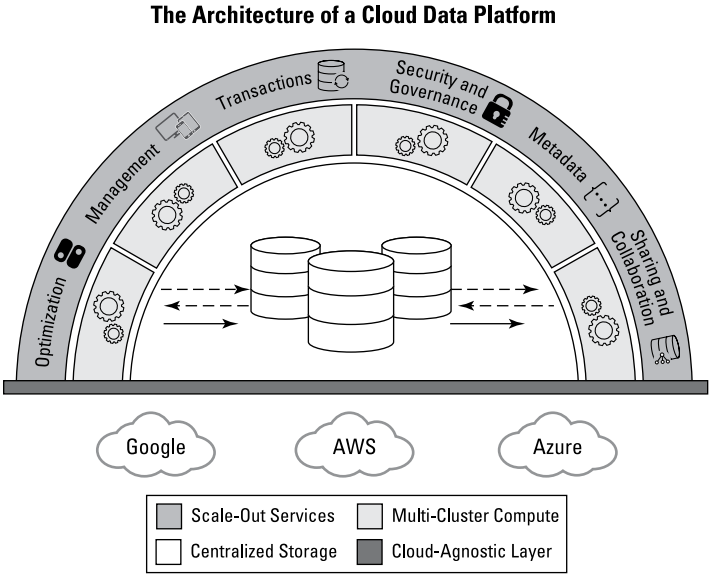


FIGURE 4-2: A comprehensive cloud data platform lets you access data from multiple clouds simultaneously and securely share data across clouds.

TIP

Your data platform should be "cloud agnostic" to provide a consistent layer of services to each cloud region and all cloud infrastructure. Look for a data platform that supports multiple public cloud environments and offers cross-cloud support. This will allow your analytic teams to store and analyze data across multiple regions and clouds.

Understanding why a cloud data platform is necessary

If you are going to put your data in an object store from a cloud storage vendor, you might wonder why you can't just use those vendors' tools for data ingestion, data governance, metadata management, and other functions. Do you even need a modern cloud data platform?

The answer is yes, mainly because a cloud data platform should allow you to work with all of these vendors' clouds in a similar fashion. It should unify workloads and resolve differences among cloud configurations by creating a common code base that spans the largest public clouds, and a common layer of services that lends cohesion to your analytic efforts. For example, a common metadata layer should ensure all users obtain consistent results, and all workloads enable consistent outcomes, no matter which public cloud resources they use. The cloud data platform should securely govern data stored in multiple clouds and across multiple regions worldwide.

Sharing data efficiently

A cloud data platform should enable organizations to easily share slices of their data and receive shared data in a secure and governed way — without requiring copying or moving data to keep it current, even when it is housed in multiple public clouds. This can help you avoid data movement and minimize the use of constant update procedures to keep data current.

Modern data sharing involves simply granting access to live, governed, read-only data by pointing at its original location. When you use the type of granular-level access control mechanisms described in Chapter 2, your data can be safely shared rather than copied, and no additional cloud storage is required. With this advanced architecture, data providers can easily and securely publish data for instant discovery, query, and enrichment by data consumers, as shown in Figure 4-3.

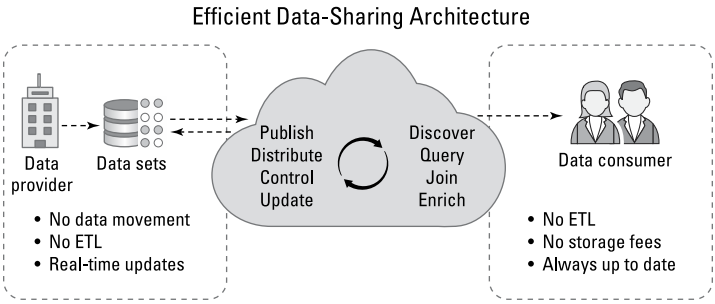


FIGURE 4-3: An efficient architecture for near real-time data sharing.

Modern cloud data platforms are designed to facilitate this type of secure data sharing. Your organization maintains one copy of

data and provides governed access to it by authorized users, both internal and external, and delivers updates to that data in near real time.


TIP

To enable advanced use cases, look for a cloud data platform that can share data on a macro level in the form of a *data exchange:* a data marketplace where customers can acquire third-party data, combine it with their own data, and share data assets on the data marketplace.

Some cloud data platforms enable customers to create their own data exchanges. This can include an organization, its suppliers and other business partners, and even its customers. Data providers incur almost no cost to share data. Data consumers pay only for compute resources, not storage, to analyze shared data. A data exchange enabled by a modern cloud data platform can minimize the cost, headache, and risk of traditional data sharing methods necessary to do business and open up new market opportunities based on previously unobtainable insights.

Sharing data shouldn't involve copying and moving data. Rather than physically transferring data to internal or external consumers, use a cloud data platform that can enable read-only access to some or all of the data set via Structured Query Language (SQL). This empowers authorized members of the cloud ecosystem to instantly access live, governed versions of the data.

Going Deeper with Visualization

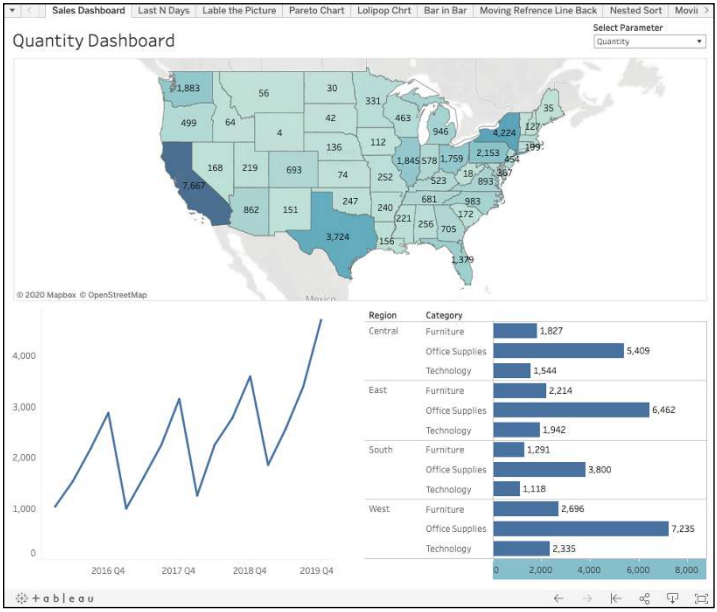
In the early years of business intelligence, data was primarily presented in textual reports and spreadsheets in relatively rigid tabular formats. Report-writing tools allowed users to specify the number of fields, width of columns, and basic layouts of each report.

Although these traditional BI tools are still popular, especially for financial reporting, many organizations now use visualization technology to make their data more accessible, approachable, and understandable. These solutions help discover insights and associations that don't jump out from a textual or numeric display. In the background, these tools generate SQL statements and database queries to retrieve results and present them on a screen via colorful charts and graphs.

Today's data visualization tools are becoming progressively easier to use and more capable of understanding what you need. For example, with some tools, you don't have to think in terms of creating a line chart or a bar chart. You simply indicate that you want to see a certain type of data, and the tool displays it for you in the most appropriate way.

This interactive approach enables a natural analytic "flow" as you iteratively discern patterns and trends in the data. You don't have to choose what type of chart or graph you want. You simply interact with your data, and the system figures out how to visualize it (see Figure 4-4). For example, if you want to see sales data over time, the system displays the data in a line chart. If you want to look at the geographic distribution of your supply chain, the data is visualized on a digital map. If you are correlating sales activity and profit margins, the tool produces a scatter plot.

Interactive Data Visualization



Source: Tableau Public

FIGURE 4-4: A sample sales dashboard that displays total sales by quarter, region, and category.

Embedding Analytics into Applications

Embedded analytics (or embedded BI) involves adding features normally associated with BI software — such as dashboard reporting, data visualization, and analytic tools — to existing applications that have limited or no analytics capabilities. This technique furthers the case for data democratization by making it easy for users to access insights in the context of the applications they use every day. In an article from McKinsey & Company titled "Breaking away: The secrets to scaling analytics," the authors suggest that presenting this type of contextual insight can be an effective way to "bridge the last mile of analytics" by delivering the right insights to the right people at the right time to drive better business outcomes.

Many organizations use embedded BI apps to share insights with customers via externally facing portals, dashboards, and reports. For example, in addition to sending you a list of your checking and credit card transactions, your bank might periodically send you an aggregated summary of all purchases organized into categories and presented visually so you can analyze your spending habits.

Some data platforms include cloud-native tools that help you develop and deploy these embedded applications, including resources for development, iteration, testing, and quality assurance (QA) activities. Connections to popular languages, tools, and utilities such as Python, Java, and SQL make it easy to build analytic applications and push them to users.

Automating Integration Tasks

One of the first things you have to figure out when building a cloud data analytics practice is how you will ingest data. Traditionally, software developers created data pipelines by hand-coding application programming interfaces (APIs), often with the help of integrated development environments (IDEs) and programming languages. Most of today's data engineers prefer to use a combination of off-the-shelf integration tools and orchestration tools to create integration logic, cleanse data, and schedule and orchestrate data pipelines.

Data integration tools allow data engineers to connect to data sources directly instead of coding APIs. *Data orchestration tools* help structure pipeline tasks such as scheduling jobs, executing

workflows, and coordinating dependencies among tasks. Most of these tools also offer self-service workbenches that allow you to select data sources, destinations, and other variables via a point-and-click environment. This makes it relatively easy to visually create data pipelines and craft the necessary software interfaces.

Some BI vendors have created direct interfaces to popular applications, such as vendor-specific connectors to CRM systems and enterprise resource planning (ERP) platforms. In some cases, they formulate three-way partnerships with a cloud data platform vendor to further simplify storing and maintaining analytic data. These integrations may include a live query engine you can use right out of the box, without having to set up data pipelines or replicate the data. You can instantly leverage all the services and resources offered by the cloud data platform vendor, such as always-on encryption and elastic storage and compute options. You immediately gain powerful analytics targeted to a specific application or business domain.

Most organizations create data pipelines that automate the transfer of data into a centralized repository such as a cloud data platform. This allows the data to be stored, managed, and transformed into specific formats necessary for various types of analysis (see Figure 4-5). To maximize your options, you need a self-service platform that can ingest both real-time events and batch data, along with flexible options for creating data pipelines that automate these data ingestion processes.

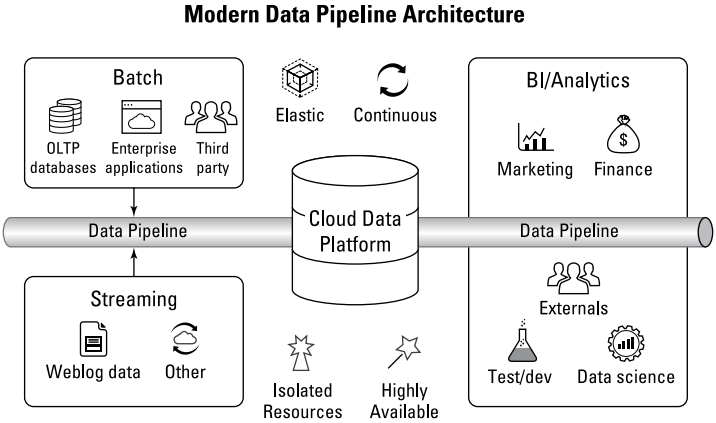


FIGURE 4-5: A modern data pipeline should support batch and streaming data, anchored by a cloud data platform as the hub for BI, analytics, and data science.

REMEMBER

Whether data arises from an enterprise application, a website, or a series of Internet of Things (IoT) sensors, data engineers must figure out how to connect to those data sources, collect the data in reliable ways, and ingest it into a cloud data platform, with the goal of making it accessible and useful to the user community.

Extracting, Transforming, and Loading Data

Modern data integration workloads are enhanced by leveraging the processing power of cloud databases and cloud data platforms that can be scaled at will. To take advantage of these cloud resources, you can design data pipelines to *extract* and *load* data and then *transform* it later once it reaches the destination, a cycle known as *ELT.* For example, data scientists commonly load data into a data lake and then combine it with another data source, or use it to train predictive models. Maintaining the data in a raw state allows them to keep their options open: They can transform the data for specific purposes once they know exactly how it will be used. This approach leverages the power of modern data processing engines and cuts down on data movement.

Traditional data pipelines use extract, transform, and load (ETL) processes to ingest data for analysis. This is a viable approach for relational data that must maintain a tabular structure. Modern data pipelines are designed to *extract* and *load* the data first and then *transform* it later (ELT) once it lands in the cloud data platform. This is a better method for semi-structured data that must be maintained in its raw or native form until specific analytic use cases are devised.

Standardizing on ANSI SQL

Many types of programming languages and query engines have arisen over the years, but SQL remains the most popular one for powering analytics and data visualization tools. It is a broadly accepted language for data query, data manipulation (insert, update, and delete), data definition (schema creation and modification), and data access control. The nonprofit organization ANSI maintains standards for SQL across technologies and geographies.

An effective cloud data analytics ecosystem should empower all users to work with all data — structured, semi-structured, and unstructured — via SQL technologies and tools. Although SQL is not the only means of building, maintaining, and querying data, your cloud data platform should allow you to ingest and query data using standard SQL and, ideally, offer a SQL database engine architected to rapidly process petabyte-scale data sets.

Taking Full Advantage of the Cloud

With a properly architected cloud data platform, all your analytic workloads can take advantage of the cloud. You can store all your data in one place, in its raw form, regardless of format, and deliver analytics-ready data to the people who need it. The platform should provide convenient access to your data and improve the speed at which you can ingest, transform, and share it across your organization and beyond. The true value of such an approach can be summed up in one word: *cohesion.* All parts of the solution fit together, including the central repository where data is stored. All data is accessible to all workloads, and those workloads leverage a common fabric of cloud services for security, identity management, transaction management, and other functions.

CASE STUDY

DEMOCRATIZING DATA ACCESS

With help from Kiva, more than 3 million borrowers from 81 countries have secured more than $1 billion in crowdfunded micro-loans. This global momentum is driven by a lofty charter: to alleviate poverty by connecting entrepreneurs with the funds they need to launch small businesses.

To accelerate lending and ensure accurate recordkeeping, Kiva ingests and analyzes large amounts of loan and relationship data into its cloud data platform, including loan data from its back-office systems, customer data from Salesforce, and additional data sets such as event data. Analysts use Looker BI tools to rapidly analyze the data and generate reports. These insights help to boost compliance by standardizing borrower verifications, loan postings, loan disbursements, and payment collections.

(continued)

(continued)

Democratizing access to the data allows many Kiva teams to work in concert by accessing a single source of truth to expand the lending community. For example, the product team shares repayment information to increase lender confidence. Customer support teams aggregate data about case interactions to elevate satisfaction with lenders, field partners, and borrowers. The finance team uses loan materialization data to help grant providers understand the cost of lending. The marketing team analyzes Salesforce data to identify top lenders and keep them engaged.

The elasticity of Kiva's cloud data platform can easily accommodate an expanding appetite for data analytics without compromising query performance. A multi-cluster shared data architecture eliminates resource contention and makes data democratization a reality, providing a scalable analytic platform for new endeavors. For example, Kiva plans to use geographic data to expand lending in underserved regions, behavior data to design new lending products and social impact programs, and data science algorithms to maximize insights from its burgeoning data sets.

IN THIS CHAPTER

- Creating a data-driven culture
- Stratifying users
- Building a business intelligence competency center
- Gaining executive guidance
- Fostering the "network effect"

Chapter 5
Driving Organizational Transformation and Alignment

Chapter 4 discusses how organizations can infuse analytics into individual activities and business processes. This chapter focuses on the people and processes that make such infusion possible.

Building Data Literacy

Successful companies understand the transformative power of data. They know how to capture, combine, and visualize data to uncover new insights. They have learned how to draw conclusions from their data based on facts rather than conjecture. These organizations have the right technology and the right people and processes to put that technology to work. They have established a culture of analytic literacy that informs the important decisions people make every day.

What are they doing right? In most cases, success is a byproduct of employing good technology *and* establishing a data-driven culture.

In a 2019 executive survey conducted by NewVantage Partners on big data and AI, executives at more than 60 percent of Fortune 1000 corporations admitted they were struggling with data-driven business transformation. Although 62 percent reported measurable results from their big data and AI investments, less than a third said they have created a data-driven organization (31 percent) or have forged a data-driven culture (28.3 percent). Nearly all of them admitted that user adoption issues stem from cultural challenges rather than from their technologies. "If companies hope to transform, they must begin to address the cultural obstacles," the report's authors advised.

REMEMBER

Simply aspiring to be data-driven is not enough. To truly be driven by data, companies must develop cultures in which this mindset can flourish. This shift in attitude begins at the top.

Gaining Executive Guidance

Gaining data literacy requires top-down prioritization as well as bottom-up enablement. Frontline workers must be empowered to act on data, but they need clear guidance on which data to use for which purposes. This is one reason solid leadership is key to the success of analytics initiatives. Someone has to own the effort, help it gain traction, and set standards to keep everybody on the same page.

Many firms hire chief data officers (CDOs) and chief analytics officers (CAOs) for this purpose. These executives help set the vision, secure funding, and drive the effort. They know that data democratization entails more than just rethinking how an organization manages, distributes, and consumes data. It also entails a dramatic shift in the attitudes people have about data. Data literacy and data-informed decision-making should be embedded into company values and modeled by leaders. It's more than just a business strategy: It affects company culture.

Building a Center of Excellence

Establishing an analytics center of excellence (CoE) can streamline and consolidate your analytics effort, especially if experts who comprise your CoE know your business and are well-acquainted

with your data sources. By combining business, IT, and analytics skills, the CoE establishes ownership for the analytics solution and helps roll out a user-adoption plan. Common tasks of the CoE include selecting the capabilities and tools, establishing standards for defining key performance indicators (KPIs) and sharing data sources, managing business intelligence (BI) projects, optimizing funding, developing skills, and training users.

As you build a CoE and identify your core analytics team, pay attention to actual usage. Foster the progression of analytics from canned dashboards and standard reports to analytic self-sufficiency. Then monitor what works and what doesn't. You may realize your staff members don't know what data sources are available or where to look for data definitions and KPIs. If that is the case, your CoE should spend more time developing training materials and best practices.

Stratifying Analytic Users

Different types of companies have different analytic maturity levels and varying levels of comfort with analytic tools. For example, professional analysts can manipulate data in its raw form. They're comfortable creating data models, joining tables, and imposing a sensible structure on a data set. They're familiar with using SQL to create and issue queries. Data scientists can build machine learning models and create predictive analytics apps. Data engineers know how to create data pipelines to populate databases and refresh analytic apps at periodic intervals. These skilled professionals may represent only 1 or 2 percent of the entire organization.

To extend the capabilities of these technical elites, identify and nurture power users — technically inclined individuals who will dive into a new tool and become its champions. Teach them to use point-and-click data visualization tools to create parameterized reports to explore the critical business factors they are investigating and self-service management dashboards for individual business functions. Train them to query, explore, and visualize data so they can spread analytics widely within their respective domains. This extends data-driven decision-making capabilities to that broad swath of the workforce interested only in accessing a report or dashboard once per week or once per month.

UNDERSTANDING THE USER COMMUNITY

A complete strategy for cloud data analytics must consider the entire community of users.

- **Application developers** use a cloud data platform to build analytic applications and generate reports for managers and executives.
- **Business users** rely on self-service BI apps that include graphical, point-and-click methods for visualizing and interacting with the data.
- **Business analysts** run queries against live data. They are comfortable writing SQL code to query data directly. They understand how to set up live connections to the database, utilize cloud services, and scale cloud data analytics to support their work.
- **Data scientists** use statistical analysis, data exploration, data mining, and modeling to build algorithms and deploy them within predictive analytic applications. They amass large amounts of data to create and train machine learning models that supply insights to applications and users.
- **Citizen data scientists** perform advanced analytics using self-service statistical and data exploration tools.
- **Power users** are technically inclined individuals who are comfortable pursuing ad hoc analytics and free-form data exploration on behalf of their departments.

Building a Core Team

Your initial team will likely include a data engineer to marshal your data, and a data analyst to serve as the data steward and build the first set of reports and dashboards. You might enlist help from one or more business users who can use prebuilt dashboards and reports to share data with others. If you have legacy databases to migrate to the cloud, a database administrator (DBA) who knows those databases is invaluable.

As the analytic practice matures, you may add app developers, data analysts, data engineers, and finally, a data scientist with experience building machine learning models or a citizen data scientist proficient with an automated machine learning (AutoML) tool.

Whether you are just starting your journey or want to advance your capabilities, enlist team members who can help you achieve your goals with cloud data analytics.

Sharing Control Between Business and IT

The IT department used to control all corporate data. However, as organizations adopted software-as-a-service (SaaS) and cloud apps, many lines of business have taken charge of their data and, in some cases, subscribed to their own analytic tools. Departmental self-sufficiency represents a positive trend. However, as data proliferates, along with the need to combine that data with lots of other sources, it has become clear that IT needs to remain in the loop, especially to oversee data governance and security.

REMEMBER

As you add more data sets, you are also introducing more "points of failure" where data can potentially be corrupted, duplicated, and so forth. It helps to have a team of specialists who are empowered to support the organization in providing accurate, timely, and complete data sets. Smart companies retain experts to help with many aspects of this process. For example, data architects build data catalogs and define metadata constructs. DBAs manage cloud and on-premises data sources. Data engineers construct data pipelines. Data stewards cleanse the data. In a small startup organization, one or two people might be responsible for all of these tasks. As a company grows and becomes more specialized, it can hire more people, all with a common goal: create efficiencies, reduce latencies, and allow more people to access accurate, holistic data safely.

Moving Up the Maturity Curve

Data is too complex for most people to readily understand and use. It comes in many different forms, and it arises from many different places. As a result, various departments may have trouble agreeing on which data should drive a particular decision. For example, do you gauge customer sentiment by looking at the customer service logs in the customer relationship management (CRM) system, or by scanning consumer posts on a social media network? As noted in a May 2020 article in *Harvard Business Review* titled "Is Your Business Masquerading as Data Driven," the goal is outcomes, not ownership.

MAXIMIZING OPERATIONAL EFFICIENCY

CASE STUDY

Although many workers can benefit from cloud data analytics, different types of people have different requirements and skill levels. Establishing a culture of analytic literacy involves developing a collaborative partnership between IT and the business units to enable self-service access to trustworthy data for everybody.

Consider Verve Wireless, a mobile marketing company that collects and analyzes signals from mobile devices to gain insight into mobile users, which it uses to help advertisers deliver relevant marketing experiences. Verve standardized on a cloud data platform that allows a diverse team of analysts and business users to load data quickly and pose many types of queries.

For example, account managers use a point-and-click interface to access daily and hourly KPIs. These visual displays allow them to keep tabs on how their campaigns perform and fine-tune their messaging for each audience. The business analytics team examines location data to identify new markets and studies mobile device attributes to identify relevant audiences. The business operations team monitors this data to ensure campaigns launch on time. Finally, the data operations team generates standard and custom reports to deliver additional insights to customers.

Thanks to the cloud data platform, Verve can ingest data once and democratize access to it without replicating or synchronizing copies. Query activities are more streamlined, whether performed via application programming interfaces (APIs) to other applications or through web and mobile dashboards. This flexible foundation has empowered many different types of users to contribute to the company's success.

Follow the advice in this chapter to help your organization move up the analytic maturity curve. Incorporate data into all your meetings, perhaps anchoring important discussions with a dashboard or report. Identify power users and teach them how to create their own analytics. The Holy Grail is to foster the "network effect," where data breeds more data and escalating analytic activities yield greater business value.

IT can encourage this trend by putting the right data and analytic processes in place. This typically includes creating a data catalog, curating the data, processing the data, and building data pipelines so analysts don't have to manage raw data. It also includes helping the user community learn about the data assets, connect to those assets, and start building from there. Keep analytics moving forward by producing a robust data set that yields useful business insights. Encourage simple proofs of concept rather than large-scale development endeavors. Operationalize the work of experts by building KPIs, dashboards, and parameterized reports. Encourage basic querying as a foundation for gradually moving to more sophisticated use cases. If you empower people with powerful tools and data, along with guidance and governance, they will use those analytic assets to make good decisions.

Adhering to Ethical Practices

As you progress with your analytic initiatives, encourage the ethical use of data and be on the lookout for hidden biases, especially in data science models. Many facial recognition tools that use machine learning, for example, have known racial biases because the model learned on pictures derived from a limited data source. Respect all pertinent regulations governing the collection, analysis, and sharing of consumer data. For guidance, you could follow the code of ethics developed by the Digital Analytics Association, a volunteer organization founded to educate organizations and Internet users about proper data collection and utilization practices.

AN ANALYTICS CODE OF ETHICS

As you build your data analytics team and capabilities, it's important to keep the following in mind:

Privacy: Always hold consumer data in the highest regard and do everything in your power to keep personally identifiable consumer data safe, secure, and private.

Transparency: Encourage full disclosure of consumer data collection practices and encourage communication of how that data will be used in clear and understandable language.

(continued)

(continued)

Consumer control: Inform and empower consumers to opt out of data collection practices, and document ways to do this.

Education: Educate users about the types of data collected and the potential risks to consumers associated with those data sets.

Accountability: Act as a steward of customer data and uphold the consumer's right to privacy as governed by applicable laws and regulations.

Considering the Future of Analytics

Success with analytics begins with widespread data literacy. To spread decision-making capabilities to casual business users, the analytic products must become easier to use. Some of this change will come in the form of natural language interfaces and AI techniques that help people interact with data in more intuitive ways. More questions will be answered by machines without human intervention. Many business processes will be digitized. Data visualization tools will allow people to think in terms that the data dictates as they answer questions in real time, or close to it.

Digital transformation initiatives succeed by getting data into the hands of people who need it as quickly as possible. The cloud allows you to keep your data and your analytics in close proximity. Cloud data platforms can make it easy to load, store, combine, and use all your data and access it via many types of analytic tools. Users don't have to think about where the data comes from or how much storage space they have. They can analyze large volumes of data to uncover patterns and insights by visualizing data in new ways.

"Business adoption of big data and AI initiatives must be viewed through a long-term lens — as a process and a journey," concluded the NewVantage Partners report. "Firms need to adopt a long-term approach, focusing on the complex cultural challenges as a starting point."

IN THIS CHAPTER

- Reviewing your current environment
- Enlisting team members
- Building a solid foundation
- Migrating data
- Launching a pilot project
- Expanding analytics throughout the enterprise

Chapter 6
Six Steps for Getting Started with Cloud Data Analytics

This book explains how to establish an analytics practice that simplifies access to all types of data as a foundation for organization-wide decision-making. It reveals how to use the cloud to broaden the scope of your analytic endeavors. And it demonstrates how you can move beyond tedious data management chores and start focusing on delivering great experiences with your data. This chapter offers six steps for getting started with cloud data analytics.

Step 1: Review Your Current State

Assess the current state of analytics at your organization by taking an inventory of your technical capabilities. Ask yourself

these questions as you determine whether to reuse, repurpose, or replace your existing assets:

- **Analytic assets:** Is your data in the cloud or on premises? How much data do you have, and from what sources? What data models and structures are in place?
- **Data assets:** Is your analytical data siloed or consolidated? Is all your data complete and accessible? Can you quickly and securely share it across your organization and with partners, suppliers, vendors, and customers?
- **Infrastructure:** Do you have an IT infrastructure that will allow you to scale your analytic efforts easily over the next two, three, or five years? Can your existing systems easily scale as you add more data, more queries, and more users? Is scalability manual or automated? Do you experience contention between storage and compute resources?
- **Data pipelines:** How robust are your data pipelines? Where are you doing your data transformations? What ingestion tools are you using? Can you accommodate a rising volume of data from a growing number of sources? Can you work with batch and streaming data?
- **IT administration:** What is the cost of maintaining analytics in your organization? Do you need to provision new infrastructure to handle occasional spikes in usage? How many people do you need to maintain your data systems?
- **Cloud resources:** What clouds are your applications built on? Do these cloud vendors make it easy to add new business intelligence tools, data science tools, and data sources, and use them all against a single source of truth?

Step 2: Enlist Your Team

Selecting and implementing the right technology is only part of what makes your analytic initiatives successful. You also need the right people and processes to guide the effort. Review the roles and responsibilities detailed in Chapter 5. Think about the types of resources you will need — from application developers and data scientists to data stewards and business analysts. Find an executive sponsor to champion the effort, and retain IT resources

to build data catalogs, define metadata, construct data pipelines, cleanse data, and assist with governance. Investigate training options from your technology vendors to empower users to prepare, explore, analyze, and model their data. Identify product owners to oversee data quality, data security, curation, lineage, and other data management functions.

Step 3: Lay a Solid Foundation

Having all your data in one unified platform opens doors to new opportunities, especially when that data is stored and managed in a consistent way, and when you can use the near-infinite capacity of the cloud to scale each workload easily and independently from one another. A comprehensive cloud data platform allows you to store data easily in its raw forms, enable immediate exploration of that data, and support a broad range of concurrent analytic use cases.

That's the starting point for success.

Use a platform with an architecture that provides the extensibility to leverage the data, tools, and capabilities you already have in place while incorporating the modern tools, processes, and procedures described in this book.

Step 4: Migrate Data

Identify the data sources you plan to analyze. Do you have historical data sets you would like to migrate to the cloud data platform? If so, you will likely want to set up a one-time transfer of this historical information. If you plan to refresh that data as new transactions occur, then you will want to establish a pipeline that can handle continuous updates.

Do you plan to store data in a public object store, such as Amazon S3, Microsoft Azure, or Google Cloud Platform? If so, make sure your data platform supports a multi-cloud architecture to maximize deployment options and offer resiliency. Consolidating data into a single source of truth, whether in a single location or across multiple cloud repositories, can provide a single, unified experience across multiple clouds and regions when accessing, analyzing, and sharing data throughout a broad analytic ecosystem.

Step 5: Launch a Pilot Project

Cloud services allow you to start small and dynamically expand your analytic ecosystem. Where do you start? Survey your organization to understand the demand for data and analytics, and then identify a manageable use case as a pilot — ideally a project that will immediately impact your business. Look for projects with executive visibility and a demonstrable return on investment (ROI). Ask yourself which users and applications will most likely benefit from cloud data analytics, and then appeal to the users and managers in these departments. Start with one project and move on to another. Help the business community conduct self-service analytics through dashboards, portals, scorecards, and other user-friendly interfaces. Then look for power users who can take it a step further by visualizing data through business intelligence (BI) tools, creating custom reports, and sharing the results with the user community.

Step 6: Prepare for Growth

Success is contagious. The results from one analytic initiative may spawn an interest in several more. Think about how you can serve your organization's basic BI and reporting needs while positioning it for advanced analytics and data science initiatives, as described in Chapter 3. Survey line-of-business managers to understand the potential of analytics in each domain. Help them address backlogs for new reports, BI applications, and ultimately predictive and prescriptive analytics. Add more technologies to your stack and deploy modern and secure data sharing and data exchange capabilities from your cloud data platform to utilize all your private and public data. And, acquire new data sets, with a focus on streamlining your business, better serving your customers, and opening up new revenue opportunities. Choose a cloud data platform that will also help you take advantage of new data, new technologies, and new capabilities as they emerge.

Standardize on a platform that can sustain rapid growth as you expand your user base and take on new initiatives. Always be cognizant of your ethical responsibilities: Respect consumer data and be alert to hidden biases in your algorithms. Consider the long-term impact of analytics on your organization and identify the people, processes, and technologies that will allow you to succeed.

Easily and securely unify, analyze, and share governed data at any scale

Your analytics team may not yet exist. Or, it needs to evolve to meet the demands of the modern enterprise. There's also the fire hose of data your organization generates but that languishes in a multitude of data silos. To remain competitive, you must also consider the deeper insights available from acquiring and integrating third-party data. This book reveals how you can build a modern data analytics strategy to easily and securely unify, analyze, and share governed data at any scale. Read on to understand what's possible for your organization with modern cloud data analytics.

Inside...

- The urgency for modern analytics
- Unifying a multitude of data silos
- Establishing a cloud-built architecture
- Grasping the potential of data science
- Driving organizational transformation and alignment
- Building a modern analytics practice
- Real-world data analytics case studies

David Baum is a freelance business writer specializing in science and technology.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-78175-2
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.